Non-consumptive research
RIT Reading Group (Feb 23, 2017)
Non-consumptive research of text corpora

- HathiTrust Research Center
- JSTOR Data For Research
- More at guides.lib.berkeley.edu/text-mining
HathiTrust.org

- Shared digital library
- Access, preservation, and storage for materials held by member libraries
- 14 million digitized volumes:
  - 7 million books
  - 725,000+ US federal gov docs
  - 350,000+ serials
HTRC (analytics.hathitrust.org)

- Provides computational research access to the HathiTrust Digital Library
- For non-profit and educational uses only (.edu emails)
- Collaboration between Indiana University and the University of Illinois
HTRC modes of access

- Portal (workset builder, pre-canned algorithms)
- Extracted Features (ngrams)
- Data capsule (secure computing environment)
- Direct researcher requests to HT
  - Non-Google public domain texts (550k)
  - Google public domain texts (4.8m) via Business Contracts Office
  - Non-public domain (?)
HTRC data capsule

- Secure VM for analysis beyond extracted features
- Restricts how and when products created leave the capsule
HTRC extracted features

1. Find volume IDs for texts of interest
2. Install HTRC Feature Reader ([in GitHub](https://github.com)) and rsync
3. Use rsync to load volume into HT Feature Reader dataframe:
   
   ```bash
   rsynchtid2rsync mdp.39015054134096 | rsync -azv --files-from=- data.sharc.hathitrust.org::features/local-folder/
   ```
JSTOR dfr (dfr.jstor.org/)

- Access to over 9 million scholarly articles in JSTOR.
- Build and download collections of N-grams for up to 1,000 documents at a time.
- Write to request access to larger sets.
- Software to pull output into R: https://github.com/benmarwick/JSTORr